Preference Inference Through Rescaling Preference Learning

Nic Wilson and Mojtaba Montazery
Insight Centre for Data Analytics
School of Computer Science and IT
University College Cork, Ireland
{nic.wilson,mojtaba.montazery}@insight-centre.org

Abstract

One approach to preference learning, based on linear support vector machines, involves choosing a weight vector whose associated hyperplane has maximum margin with respect to an input set of preference vectors, and using this to compare feature vectors. However, as is well known, the result can be sensitive to how each feature is scaled, so that rescaling can lead to an essentially different vector. This gives rise to a set of possible weight vectors—which we call the rescale-optimal ones—considering all possible rescalings. From this set one can define a more cautious preference relation, in which one vector is preferred to another if it is preferred for all rescale-optimal weight vectors. In this paper, we analyse which vectors are rescale-optimal, and when there is a unique rescale-optimal vector, and we consider how to compute the induced preference relation.

1 Introduction

In many contemporary application domains, for example, information retrieval from large databases or the web, or planning in complex domains, the user has little knowledge about the set of possible solutions or feasible items, and what they generally seek is the best that's out there. But since the user does not know what is the best achievable, they typically cannot characterize it or its properties specifically [Brafman, 2008]. So, it is desirable for the system to learn the user's preferences over alternative choices (that is, documents, movies, products and so on) [Brafman and Domshlak, 2009].

Generally, a preference learning task consists of some set of items for which preferences are known, and the task is to learn a function which predicts preferences for a new set of items. An established approach to modeling preferences makes use of the concept of a *utility function*. Such a function assigns an abstract degree of utility to each alternative under consideration [Fürnkranz and Hüllermeier, 2010]. Learning a utility function from a given set of training data could be viewed from a machine learning perspective. However, training data is not necessarily given in the form of input/output pairs, but may consist of qualitative feedback in the form of pairwise comparisons, stating that one alternative is preferred to another one and therefore has a higher utility degree. Support Vector Machine (SVM) approaches [Burges, 1998] are popular in machine learning, and have inspired the development of several methods for preference learning, such as OrderSVM [Kazawa *et al.*, 2005], SVOR [Herbrich *et al.*, 1999] and SVMRank [Joachims, 2002]. Essentially, SVM-based methods are built under this assumption that the utility function is a linear weighted sum of the features. Despite the fact that a linear structure for the preference function may sound too restrictive, incorporating the *kernel* trick [Aizerman *et al.*, 1964] alleviates this by providing more flexibility, to model non-linear functions as well.

Feature spaces normalization (scaling) is an essential requirement for any SVM-based method because they are not invariant to the scale of their input feature spaces: multiplying a feature dimension by a fixed constant > 1 gives that dimension more weight in the value of the SVM objective function and, therefore, in the choice of the weight vector in the preferences function [Stolcke *et al.*, 2008; Ben-Hur and Weston, 2010]. If we base the scalings on the input instances, then it can make the induced preference relation sometimes highly sensitive to precisely which instances are received. There can thus be subjective, and even rather arbitrary, choices in the scaling of the feature spaces; different ways lead to different preference relations. This suggests defining a more cautious preference relation, consisting of all pairs that are inferred for all choices of scalings.

Thus, one alternative is preferred to another if it is preferred for all *rescale-optimal* weight vectors, where the rescale-optimal vectors are those that can be made optimal for some rescaling. This is a form of preference inference; a related form is when we only keep preferences that are supported by all compatible weight vectors, which corresponds to the kind of preference inference considered in [Marinescu *et al.*, 2013]. Other forms of preference inference, based on more qualitative, lexicographic, models are considered in [Trabelsi *et al.*, 2011; Wilson, 2014; Wilson *et al.*, 2015]. Other preference reasoning techniques based on a family of utility functions include e.g., [Greco *et al.*, 2010].

In this paper we analyse the new preference relation, deriving results regarding rescale-optimality that entail when scaling makes a difference, and that lead to a characterisation that

allows computation of preference.

Section 2 defines the preference relation, and the notion of rescale-optimality, and gives some basic properties. It can happen that rescaling makes no difference; we show how to determine this in Section 3. In Sections 4 and 5 we give characterisations for rescale-optimality, that lead to a way of computing the relation, which we test out with benchmarks derived from a real ride-sharing dataset in Section 6.

2 Rescaled Maximum Margin Preference Learning

In this section we define, and give some properties of, a preference relation based on rescaling maximum margin preference learning.

2.1 Some Terminology

We first list some terminology that we'll be using throughout the paper. Consider arbitrary $u, v, w \in \mathbb{R}^n$. We say that v is non-zero if v is not equal to the zero vector $\mathbf{0} = (0, \ldots, 0) \in \mathbb{R}^n$. Also, v is said to be strictly positive if $v(j) > 0$ for all $j = 1, \ldots, n$; let $\mathbb{R}^n_+$ be the set of strictly positive elements in $\mathbb{R}^n$.

The dot product $u \cdot v$ is equal to $\sum_{j=1}^n u(j)v(j)$. The (Euclidean) norm $|u|$ of u is given by $|u|^2 = \sum_{j=1}^n (u(j))^2$, which equals $u \cdot u$. We also define $u \odot v$ to be the vector in $\mathbb{R}^n$ given by pointwise multiplication, and thus, for all $j = 1, \ldots, n$, $(u \odot v)(j) = u(j)v(j)$. Operation $\odot$ is commutative, associative and distributes over addition of vectors. An important property is that for any $u, v, w \in \mathbb{R}^n$ $(u \odot v) \cdot w = v \cdot (u \odot w)$, since they are both equal to $\sum_{j=1}^n u(j)v(j)w(j)$.

For $\Lambda \subseteq \mathbb{R}^n$, define sets Λ^*, $\Lambda^>$ and $\Lambda^{\geq 1}$ by

$$\Lambda^* = \{w \in \mathbb{R}^n : \forall \lambda \in \Lambda, w \cdot \lambda \geq 0\};$$
$$\Lambda^> = \{w \in \mathbb{R}^n : \forall \lambda \in \Lambda, w \cdot \lambda > 0\}; \text{ and}$$
$$\Lambda^{\geq 1} = \{w \in \mathbb{R}^n : \forall \lambda \in \Lambda, w \cdot \lambda \geq 1\}.$$

For finite $\Lambda \subseteq \mathbb{R}^n$, we define $co(\Lambda)$ to be the convex cone generated by Λ, i.e., the set of all vectors in $\mathbb{R}^n$ that can be written as $\sum_{v \in \Lambda} r_v v$, where r_v are arbitrary non-negative reals. Then $co(\Lambda)$ is finitely generated (by Λ). Elements of $co(\Lambda)$ are said to be *positive linear combinations of elements of* Λ. A *polyhedron* is the intersection of a finite number of closed half-spaces, so is topologically closed and convex and can be written as $\{w \in \mathbb{R}^n : \forall i \in I, w \cdot \lambda_i \geq a_i\}$, for finite I, where each $\lambda_i \in \mathbb{R}^n$ and $a_i \in \mathbb{R}$.

2.2 Maximum Margin Preference Relation

We first describe a simple linear SVM-based preference relation, based on Ranking SVM [Joachims, 2002], but only considering consistent inputs. Let Λ and Θ be finite subsets of $\mathbb{R}^n$. We call Λ, the *preference inputs*, and we call Θ the *constraints*. Each input preference $\lambda \in \Lambda$ expresses a linear restriction $\lambda \cdot w > 0$ on an unknown user weight vector $w \in \mathbb{R}^n$. For instance, if there are n-features, and the user has told us that they prefer feature vector α to β (where $\alpha, \beta \in \mathbb{R}^n$, each representing the values of the alternative over the n features), then we induce from this that $\alpha \cdot w > \beta \cdot w$, i.e., $(\alpha - \beta) \cdot w > 0$, so we include $\alpha - \beta$ in Λ. (This linear weighting assumption is less restrictive than it sounds; for instance, we could form additional features representing e.g., pairwise products of the basic features, enabling a richer representation of the utility function.) The constraints set Θ is used for placing general restrictions on w; in particular, for expressing a restriction that higher values of the jth feature are at least as good; this translates to a constraint of the form $w(j) \geq 0$, represented by including e_j in Θ, where e_j is the jth unit vector, with $e_j(j) = 1$, and $e_j(k) = 0$ for $j \neq k$.

The *feasible set* $C(\Lambda, \Theta)$ is defined to be $\Lambda^> \cap \Theta^*$. We also define $G(\Lambda, \Theta)$ to be $\Lambda^{\geq 1} \cap \Theta^*$. If Θ is empty, we may abbreviate $G(\Lambda, \Theta)$ to $G(\Lambda)$ (and similarly, for other definitions).

The margin function $\text{marg}_\Lambda : C(\Lambda, \Theta) \to \mathbb{R}$ is given by $\text{marg}_\Lambda(w) = \min_{\lambda \in \Lambda} \frac{w \cdot \lambda}{|w|}$. This is equal to the distance between hyperplane $H_w = \{\mu \in \mathbb{R}^n : \mu \cdot w = 0\}$ and the closest element of Λ to H_w. The definition implies that $\text{marg}_\Lambda(w) > 0$ for all $w \in C(\Lambda, \Theta)$, since $w \cdot \lambda > 0$ for any $w \in \Lambda^>$ and $\lambda \in \Lambda$. One might view $\frac{w \cdot \lambda}{|w|}$ as the degree to which w satisfies the preference λ, with the best w being those that satisfy each λ to the greatest degree, i.e., those that maximise $\text{marg}_\Lambda(w)$.

For $w \in \mathbb{R}^n$ we define the associated relation $>_w$ by, for $\alpha, \beta \in \mathbb{R}^n$, $\alpha >_w \beta$ if and only if $w \cdot \alpha > w \cdot \beta$. Note that for any real $r > 0$, the relation $>_{rw}$ is equal to $>_w$.

Define the preference relation $>^{mm}_{\Lambda,\Theta}$ by, for $\alpha, \beta \in \mathbb{R}^n$, $\alpha >^{mm}_{\Lambda,\Theta} \beta$ if and only if there exists w with maximum margin in $C(\Lambda, \Theta)$ such that $\alpha >_w \beta$. Theorem 1 implies that $\alpha >^{mm}_{\Lambda,\Theta} \beta$ if and only if $\alpha >_{w(\Lambda,\Theta)} \beta$, where $w(\Lambda, \Theta)$ is the element in $G(\Lambda, \Theta)$ with minimum norm.[1] Thus, $>^{mm}_{\Lambda,\Theta}$ is close to being a total order, since for any $\alpha, \beta \in \mathbb{R}^n$ we have either $\alpha >^{mm}_{\Lambda,\Theta} \beta$ or $\beta >^{mm}_{\Lambda,\Theta} \alpha$ or $w(\Lambda, \Theta) \cdot (\alpha - \beta) = 0$.

Theorem 1 *For finite $\Lambda, \Theta \subseteq \mathbb{R}^n$, if $C(\Lambda, \Theta)$ is non-empty then $G(\Lambda, \Theta)$ is non-empty and there exists a unique element $w(\Lambda, \Theta)$ in $G(\Lambda, \Theta)$ with minimum norm. Also, for $w \in C(\Lambda, \Theta)$, w maximises marg_Λ within $C(\Lambda, \Theta)$ if and only if w is a strictly positive scalar multiple of $w(\Lambda, \Theta)$, i.e., there exists $r \in \mathbb{R}$ with $r > 0$ such that $w = r w(\Lambda, \Theta)$.*

Example 1: Suppose that $\Lambda = \{(2, -1), (-1, 2), (\frac{1}{5}, \frac{1}{5})\}$ with Θ being empty. Then the feasible set $C(\Lambda, \Theta) = \Lambda^>$ which equals the set of all $w \in \mathbb{R}^n$ such that $2w(1) > w(2)$ and $2w(2) > w(1)$. The set $G(\Lambda, \Theta) = \Lambda^{\geq 1}$ is shown in Figure 1(a); it has two extremal points: $(3, 2)$ (corresponding to the intersection of the lines $2y - x = 1$ and $x + y = 5$) and $(2, 3)$. The point in $G(\Lambda)$ with minimum norm is $(2.5, 2.5)$. Theorem 1 implies that the elements in $C(\Lambda, \Theta)$ with maximum margin are w with $w(1) = w(2)$. For instance, if $w = (1, 1)$ then $\text{marg}_\Lambda(w) = \frac{1}{\sqrt{2}} \min(1, 1, \frac{2}{5}) = \frac{\sqrt{2}}{5}$. The relation $>^{mm}_{\Lambda,\Theta}$ equals $>_{(2.5,2.5)}$, which is the same as $>_{(1,1)}$. Thus, $\alpha >^{mm}_{\Lambda,\Theta} \beta$, i.e., $\alpha - \beta >^{mm}_{\Lambda,\Theta} \mathbf{0}$, if and only

[1] Because of the space restrictions, not all the proofs could be included. See [Wilson and Montazery, 2016] for missing proofs.

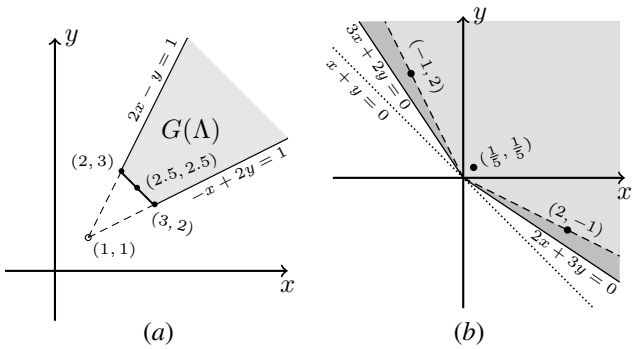

Figure 1: (a) The shaded region shows $G(\Lambda)$ when $\Lambda = \{(2,-1),(-1,2),(\frac{1}{5},\frac{1}{5})\}$, with every element of the line segment between $(3,2)$ and $(2,3)$ being rescale-optimal in $G(\Lambda)$. (b) shows the boundaries of the preferred regions for the relations $>^{mm}_{\Lambda}$, $\gg_{\Lambda}$ and $\gg^{*}_{\Lambda}$.

if $(\alpha-\beta)\cdot(1,1) > 0$, i.e., $\alpha(1)+\alpha(2)-(\beta(1)+\beta(2)) > 0$. The set of γ that are $>^{mm}_{\Lambda,\Theta}$-preferred to $\mathbf{0}$ is the region in Figure 1(b) to the right of the dotted line.

2.3 The Effect of Rescaling

Consider the effect of a rescaling τ, i.e., $\tau \in I\!R^n_+$, where an element $v \in I\!R^n$ is transformed into $v \odot \tau$. An input preference vector λ becomes $\lambda \odot \tau$, so Λ becomes $\Lambda \odot \tau = \{\lambda \odot \tau : \lambda \in \Lambda\}$, and the constraints set Θ becomes $\Theta \odot \tau$. The feasible set becomes $C(\Lambda \odot \tau, \Theta \odot \tau)$. For example, if τ is the rescaling $(1,2)$ and $\Lambda = \{(2,-1),(-1,2),(\frac{1}{5},\frac{1}{5})\}$ then $\Lambda \odot \tau = \{(2,-2),(-1,4),(\frac{1}{5},\frac{2}{5})\}$. Rescaling by τ means that each $\alpha \in I\!R^n$ becomes instead $\alpha \odot \tau$. Let us say that α *is max-margin-preferred to* β *under rescaling* τ if $\alpha \odot \tau >^{mm}_{\Lambda\odot\tau,\Theta\odot\tau} \beta \odot \tau$, i.e., if rescaled α is preferred to rescaled β under the max margin relation relation corresponding to rescaled Λ and Θ. Now, it can easily happen that α is preferred to β under one rescaling, but not under another (see e.g., the example in Section 2.5). Also, the choice of how the features are scaled relative to each other can involve somewhat arbitrary choices. It is therefore natural to consider the relation given by α being preferred to β for all rescalings $\tau \in I\!R^n_+$.

Definition 1 ($\gg_{\Lambda,\Theta}$) *We define relation $\gg_{\Lambda,\Theta}$ by, for $\alpha,\beta \in I\!R^n$, $\alpha \gg_{\Lambda,\Theta} \beta$ if and only if α is max-margin-preferred to β over all rescalings, i.e., if for all $\tau \in I\!R^n_+$, $\alpha \odot \tau >^{mm}_{\Lambda\odot\tau,\Theta\odot\tau} \beta \odot \tau$.*

Let $w^*_\tau(\Lambda,\Theta)$ (abbreviated to w^*_τ) be the element with minimum norm in $G(\Lambda \odot \tau, \Theta \odot \tau)$. Theorem 1 implies that $\alpha \odot \tau >^{mm}_{\Lambda\odot\tau,\Theta\odot\tau} \beta \odot \tau$ if and only if $\alpha \odot \tau >_{w^*_\tau} \beta \odot \tau$, which can be rewritten as $\alpha >_{w^*_\tau \odot \tau} \beta$.

Defining $\mathrm{RO}(\Lambda,\Theta)$ to be $\{w^*_\tau \odot \tau : \tau \in I\!R^n_+\}$, we have:

$$\alpha \gg_{\Lambda,\Theta} \beta \iff \text{for all } w \in \mathrm{RO}(\Lambda,\Theta),\ \alpha >_w \beta.$$

For example, it can be shown that in Figure 1(a), $\mathrm{RO}(\Lambda,\Theta)$ is the closed line segment between $(2,3)$ and $(3,2)$. We show below that $\mathrm{RO}(\Lambda,\Theta)$ is equal the set of rescale-optimal elements in $G(\Lambda,\Theta)$, defined as follows.

Definition 2 (Rescale-optimal) *For $G \subseteq I\!R^n$, and $u \in G$, let us say that u* is rescale-optimal in G *if there exists strictly positive $\tau \in I\!R^n_+$ with $|\tau \odot w| \geq |\tau \odot u|$ for all $w \in G$.*

If $\mathbf{0} \in G$ then it is the unique element that is rescale-optimal in G. For $\tau \in I\!R^n_+$, we define τ^{-1} to be the element of $I\!R^n_+$ given by $\tau^{-1}(j) = 1/\tau(j)$ for all $j \in \{1,\ldots,n\}$.

Lemma 1 *Consider any $v \in I\!R^n$ and any $\tau \in I\!R^n_+$. Then, $v \in G(\Lambda,\Theta)$ if and only if $v \odot \tau \in G(\Lambda \odot \tau^{-1}, \Theta \odot \tau^{-1})$. Also, $v = w$ minimises $|w \odot \tau|$ over $w \in G(\Lambda,\Theta)$ if and only if $v = \tau^{-1} \odot w^*_{\tau^{-1}}(\Lambda,\Theta)$.*

Proposition 1 $\mathrm{RO}(\Lambda,\Theta)$ *is the set of all rescale-optimal elements of $G(\Lambda,\Theta)$. Thus, for $\alpha,\beta \in I\!R^n$, $\alpha \gg_{\Lambda,\Theta} \beta$ if and only if $\alpha >_w \beta$ for all rescale-optimal elements of $G(\Lambda,\Theta)$, which is if and only if $\alpha - \beta \in (\mathrm{RO}(\Lambda,\Theta))^{>}$.*

Proof: Consider any $v \in I\!R^n$. Then, v is rescale-optimal in $G(\Lambda,\Theta)$ if and only if there exists $\tau \in I\!R^n_+$ such that $v = w$ minimises $|w \odot \tau|$ over $w \in G(\Lambda,\Theta)$, which, by Lemma 1, is if and only if there exists $\tau \in I\!R^n_+$ such that $v = \tau^{-1} \odot w^*_{\tau^{-1}}(\Lambda,\Theta)$, which is if and only if $v \in \mathrm{RO}(\Lambda,\Theta)$. The last part follows immediately from the definitions. □

Another natural preference relation $\gg^*_{\Lambda,\Theta}$, which is very closely related to the one explored in [Marinescu *et al.*, 2012; 2013], is given by $\alpha \gg^*_{\Lambda,\Theta} \beta$ if and only if for all $w \in C(\Lambda,\Theta)$, $w \cdot (\alpha - \beta) > 0$, i.e., iff α is preferred to β for all compatible weight vectors. This holds if and only if for all $w \in G(\Lambda,\Theta)$, $w \cdot (\alpha - \beta) > 0$. Thus, $\alpha \gg^*_{\Lambda,\Theta} \beta$ if and only if $\alpha - \beta \in (G(\Lambda,\Theta))^{>}$. (When Θ is empty, $(G(\Lambda,\Theta))^{>}$ is equal to $co(\Lambda)$, the smallest convex cone containing Λ.) This implies that if $\alpha \gg^*_{\Lambda,\Theta} \beta$ then $\alpha \gg_{\Lambda,\Theta} \beta$. The three defined preference relations are therefore nested: $\gg^*_{\Lambda,\Theta} \subseteq \gg_{\Lambda,\Theta} \subseteq >^{mm}_{\Lambda,\Theta}$. The three relations are all irreflexive and transitive, and thus strict partial orders (with $>^{mm}_{\Lambda,\Theta}$ close to being a total order). Also, if $\succ$ is any of the three relations, $\lambda \succ \mathbf{0}$, for any $\lambda \in \Lambda$, and for $\alpha,\beta,\gamma \in I\!R^n$ and $r \in I\!R$ with $r > 0$, if $\alpha \succ \beta$ then $\alpha + \gamma \succ \beta + \gamma$ and $r\alpha > r\beta$.

2.4 Pointwise Undominated Vectors

For $u \in G(\Lambda,\Theta)$, if there exists $v \in G(\Lambda,\Theta)$ such that for all j, $v(j)$ is between $u(j)$ and 0 then it is easy to see that u is not rescale-optimal. This is the idea behind being pointwise undominated.

Definition 3 (pointwise dominance) *For $u,v \in I\!R^n$, v pointwise dominates u if $u \neq v$ and for all $j \in \{1,\ldots,n\}$, either $0 \leq v(j) \leq u(j)$ or $0 \geq v(j) \geq u(j)$. u is pointwise undominated in $G \subseteq I\!R^n$ if there exists no $v \in G$ that pointwise dominates u.*

The definition easily implies that rescale-optimality implies being pointwise undominated.

Proposition 2 *Let $G \subseteq I\!R^n$. If u is rescale-optimal in G then u is pointwise undominated in G.*

2.5 Example 1 Continued

Since $(2.5, 2.5)$ has minimum norm in $G(\Lambda)$, it is rescale-optimal. The only pointwise undominated elements in $G(\Lambda)$ are $(3, 2)$, $(2, 3)$, and the line segment joining them, and, it turns out that all these are rescale-optimal. If we set $\tau = (r, 1)$ then $r \leq \sqrt{2/3}$ will lead to the point $(3, 2)$ having minimum value of $|(x, y) \odot \tau|$.

Figure 1(b) illustrates the strength of the three relations, $>^{mm}_{\Lambda}$, $\gg_{\Lambda}$ and $\gg^{*}_{\Lambda}$, by showing the regions of all $\gamma \in I\!R^2$ that are preferred to $\mathbf{0}$. For $\gg^{*}_{\Lambda}$, this is the convex cone generated by the three input vectors Λ, i.e., all positive linear combinations of them. For $\gg_{\Lambda}$, the preferred set is the intersection of the halfspaces $\{\gamma : \gamma \cdot (3, 2) > 0\}$ and $\{\gamma : \gamma \cdot (2, 3) > 0\}$. The largest preference region is associated with $>^{mm}_{\Lambda}$, with $\gamma >^{mm}_{\Lambda,\Theta} \mathbf{0}$ if and only if $\gamma(x) + \gamma(y) > 0$. Since $\gg_{\Lambda}$ is not equal to $>^{mm}_{\Lambda}$, the scaling makes a difference. E.g., $(5, 0) >^{mm}_{\Lambda} (0, 4)$, which is the same as $(5, 0)$ being preferred to $(0, 4)$ under rescaling $(1, 1)$. However, it can be shown that $(0, 4)$ is preferred to $(5, 0)$ under rescaling $\tau = (1, 2)$, with $w^{*}_{\tau} \odot \tau$ being equal to $(2, 3)$ (which minimises $|v \odot \tau^{-1}|$ over $v \in G(\Lambda)$) and $(0, 4) \cdot (2, 3) > (5, 0) \cdot (2, 3)$.

Now consider if Λ instead equals $\{(2, -1), (-1, 2)\}$. Then $G(\Lambda)$ has a single extremal point $(1, 1)$, being the intersection of the lines $2x - y = 1$ and $2y - x = 1$. Since $(1, 1)$ is the element in $G(\Lambda)$ with minimum norm, we have $w(\Lambda, \Theta) = (1, 1)$, and $(1, 1)$ is rescale-optimal. Thus, $>^{mm}_{\Lambda}$ equals $>_{(1,1)}$. In fact, $(1, 1)$ pointwise dominates every other element in $G(\Lambda)$, so, by Proposition 2, is the only rescale-optimal vector in $G(\Lambda)$. Then, $\gg_{\Lambda}$ is just $>_{(1,1)}$, so that $\alpha \gg_{\Lambda} \beta$ if and only if $\alpha(x) + \alpha(y) > \beta(x) + \beta(y)$. This example shows that allowing rescaling can sometimes make no difference. In Section 3 we show a general result that u is a unique rescale-optimal element in closed convex G if and only if u pointwise dominates every other element of G.

3 Determining Uniquely Rescale-Optimal Vectors

Here we characterise the situations when rescaling makes no difference, i.e., when there is a unique rescale-optimal vector.

For convex closed $G \subseteq I\!R^n$, and for $\tau \in I\!R^n_+$ we write $w_\tau(G)$ for the unique vector $w \in G$ with minimum value of $|w \odot \tau|$, which makes sense because of the following result. Thus, u is rescale-optimal in convex closed G if and only if there exists $\tau \in I\!R^n_+$ such that $u = w_\tau(G)$.

Lemma 2 *Let G be a convex and (topologically) closed subset of $I\!R^n$. For each strictly positive vector $\tau \in I\!R^n_+$, there exists a unique $w \in G$ with minimum value of $|w \odot \tau|$.*

Theorem 2 below states that u is the only rescale-optimal element in convex closed G if and only if u pointwise dominates every other element of G. The proof uses a pair of lemmas.

Lemma 3 *Let $u, v \in I\!R^n$. There exists $k \in \{1, \ldots, n\}$ such that $|u(k)| < |v(k)|$ if and only if there exists $\tau \in I\!R^n_+$ such that $|u \odot \tau| < |v \odot \tau|$. Thus, for all $j \in \{1, \ldots, n\}$, $|u(j)| \geq |v(j)|$ if and only if for all $\tau \in I\!R^n_+$, $|u \odot \tau| \geq |v \odot \tau|$.*

Lemma 4 *Let G be a convex subset of $I\!R^n$, and let j be any element of $\{1, \ldots, n\}$. Then either (i) there exists $w \in G$ such that $w(j) = 0$; or (ii) for all $w \in G$, $w(j) > 0$; or (iii) for all $w \in G$, $w(j) < 0$.*

Theorem 2 *Let G be a convex and closed subset of $I\!R^n$, and let u be an element of G. Then the following conditions are equivalent.*

(i) u is uniquely rescale-optimal in G, i.e., u is the unique element of G that is rescale-optimal;

(ii) for all $v \in G$, for all $j \in \{1, \ldots, n\}$, $|v(j)| \geq |u(j)|$;

(iii) u pointwise dominates every element in $G - \{u\}$.

The equivalence between (i) and (ii) is proved using Lemmas 2 and 3, and the equivalence between (ii) and (iii) follows using Lemma 4.

Corollary 1 *For finite $\Lambda, \Theta \subseteq I\!R^n$, let $G = G(\Lambda, \Theta)$. Define $y \in I\!R^n$ as follows. Choose an arbitrary element $z \in G$. For each $j \in \{1, \ldots, n\}$: If $z(j) = 0$ then define $y(j) = 0$. If $z(j) > 0$ then define $y(j) = \inf\{w(j) : w \in G, w(j) \geq 0\}$. If $z(j) < 0$ then define $y(j) = \sup\{w(j) : w \in G, w(j) \leq 0\}$. If $y \in G$ then y is uniquely rescale-optimal in G. Also, there exists a uniquely rescale-optimal element in G if and only if $y \in G$.*

Corollary 1 leads immediately to an algorithm for determining if $G(\Lambda, \Theta)$ has a uniquely rescale-optimal element, and finding it, if it exists. The algorithm involves at most $n+1$ runs of a linear programming solver, and thus determining and finding a uniquely rescale-optimal element u can be performed in polynomial time. If it succeeds in finding such a u then the induced preferences can be efficiently tested using: $\alpha \gg_{\Lambda,\Theta} \beta$ if and only if $u \cdot (\alpha - \beta) > 0$.

4 Zm-Pointwise Undominated Vectors

Proposition 2 states that being pointwise undominated is a necessary condition for being rescale-optimal. The example below shows that the two conditions are not equivalent. In this section we define a stronger version of pointwise undominated called *zm-pointwise undominated*, where 'zm' stands for *zeros-modified* (the essential difference being in the treatment of j such that $u(j) = 0$). We show that this is still a necessary condition for rescale-optimality, and is in fact equivalent to rescale-optimality (for polyhedra).

Example 2: Let $G \subseteq I\!R^2$ be given by all pairs (x, y) such that $x + y \geq 1$. It can be seen that the set of points that are pointwise undominated is $\{(x, 1 - x) : x \in [0, 1]\}$. On the other hand, the set of points that are rescale-optimal is $\{(x, 1 - x) : x \in (0, 1)\}$: neither $(1, 0)$ nor $(0, 1)$ is rescale-optimal. This is because if rescaling $\tau \in I\!R^2$ is such that $\tau(x)/\tau(y) = r$, for $r \in (0, \infty)$, then the associated rescale-optimal $w_\tau(G)$ is equal to $\frac{1}{1+r^2}(1, r^2)$, which is never equal to $(1, 0)$ or $(0, 1)$.

Definition 4 (zm-pointwise undominated) *We say that u is zm-pointwise undominated in G if for all $v \in G$, either (a) $v(j) = u(j)$ for all $j \in \{1, \ldots, n\}$ such that $u(j) \neq 0$; or (b) there exists $k \in \{1, \ldots, n\}$ such that either $0 < u(k) < v(k)$ or $0 > u(k) > v(k)$.*

It is easily shown that if u is zm-pointwise undominated in convex G then it is pointwise undominated in G. Proposition 3 below shows that being zm-pointwise undominated is a necessary condition for being rescale-optimal. The proof uses the following lemma.

Lemma 5 *Let $u, v \in \mathbb{R}^n$, with $u \neq v$, and let $\tau \in \mathbb{R}^n_+$. For $\delta \in (0,1]$ let $v_\delta = \delta v + (1-\delta)u$. Then the following hold:*

(i) For any $\delta \in \mathbb{R}$, $|v_\delta \odot \tau|^2 - |u \odot \tau|^2 = \delta^2 |(v-u) \odot \tau|^2 + 2\delta(\tau \odot \tau \odot u) \cdot (v-u)$.

(ii) $(\tau \odot \tau \odot u) \cdot (v-u) \geq 0$ if and only if for all $\delta \in (0,1]$, $|v_\delta \odot \tau| > |u \odot \tau|$.

(iii) There exists $\tau \in \mathbb{R}^n_+$ such that $(\tau \odot \tau \odot u) \cdot (v-u) \geq 0$ if and only if either (a) $v(j) = u(j)$ for all $j \in \{1, \ldots, n\}$ such that $u(j) \neq 0$; or (b) there exists $k \in \{1, \ldots, n\}$ such that either $0 < u(k) < v(k)$ or $0 > u(k) > v(k)$.

Proposition 3 *Let u be an element of convex $G \subseteq \mathbb{R}^n$. Then:*

(i) u is rescale-optimal in G if and only if there exists $\tau \in \mathbb{R}^n_+$ such that for all $v \in G$, $(\tau \odot \tau \odot u) \cdot (v-u) \geq 0$.

(ii) u is zm-pointwise undominated in G if and only if for all $v \in G$, there exists $\tau \in \mathbb{R}^n_+$ such that $(\tau \odot \tau \odot u) \cdot (v-u) \geq 0$.

(iii) If u is rescale-optimal in G then u is zm-pointwise undominated in G.

Proof: (i): Using Lemma 2, u is rescale-optimal in G if and only if there exists $\tau \in \mathbb{R}^n_+$ such that for all $v \in G - \{u\}$, $|v \odot \tau| > |u \odot \tau|$, which, since G is convex, is if and only if, there exists $\tau \in \mathbb{R}^n_+$ such that for all $v \in G - \{u\}$ and for all $\delta \in (0,1]$, $|v_\delta \odot \tau| > |u \odot \tau|$, where $v_\delta = \delta v + (1-\delta)u$. By Lemma 5(ii), this is if and only if there exists $\tau \in \mathbb{R}^n_+$ such that for all $v \in G - \{u\}$, $(\tau \odot \tau \odot u) \cdot (v-u) \geq 0$, which holds iff for all $v \in G$, $(\tau \odot \tau \odot u) \cdot (v-u) \geq 0$.

(ii) By Lemma 5(iii), u is zm-pointwise undominated in G if and only if for all $v \in G - \{u\}$, there exists $\tau \in \mathbb{R}^n_+$ such that $(\tau \odot \tau \odot u) \cdot (v-u) \geq 0$, from which (ii) follows.

(iii) follows immediately from (i) and (ii). □

Definition 5 (agreeing on signs) *For $u, v \in \mathbb{R}^n$, we say that u* and *v* agree on signs *if for all $j = 1, \ldots, n$, (i) $u(j) = 0 \iff v(j) = 0$; (ii) $u(j) > 0 \iff v(j) > 0$; and thus also: (iii) $u(j) < 0 \iff v(j) < 0$.*

Proposition 3(i) implies the following characterisation of rescale-optimality, by letting $\mu' = \tau \odot \tau \odot u$ and $\mu = \frac{\mu'}{\mu' \cdot u}$; for the converse, we use τ such that $\mu = \tau \odot \tau \odot u$, so that $\tau(j)^2 = \mu(j)/u(j)$ when $u(j) \neq 0$.

Theorem 3 *Consider any u in convex $G \subseteq \mathbb{R}^n$. If $u = \mathbf{0}$ then it is the unique rescale-optimal element of G. Otherwise, u is rescale-optimal in G if and only if there exists $\mu \in \mathbb{R}^n$ agreeing on signs with u such that $\mu \cdot u = 1$ and for all $w \in G$, $\mu \cdot w \geq 1$.*

It turns out that being zm-pointwise undominated is equivalent to being rescale-optimal, for a polyhedron. (The proof is quite technical, and makes use of classical results about convex sets, see [Wilson and Montazery, 2016].)

Theorem 4 *Let u be an element of polyhedron $G \subseteq \mathbb{R}^n$. Then, u is rescale-optimal in G if and only if u is zm-pointwise undominated in G.*

5 Computational Characterisation of Rescale-Optimal

Here we extend the characterisation of rescale-optimality given in Theorem 4, leading to a computational method for testing rescale-optimality, and thus to a method for testing if $\alpha \gg_{\Lambda,\Theta} \beta$, for $\alpha, \beta \in \mathbb{R}^n$.

5.1 Expressing Rescale-Optimality in Terms of Positive Linear Combinations

Theorem 3 implies that non-zero u is rescale-optimal in $G(\Lambda, \Theta)$ if and only if there exists a vector μ that agrees on signs with u with $\mu \cdot w \geq \mu \cdot u$ for all $w \in G$. The main result of this section is the following, that shows that μ is a positive linear combination of certain vectors in $\Lambda \cup \Theta$.

Theorem 5 *Let G be a polyhedron, which we write as $G_I = \{w \in \mathbb{R}^n : \forall i \in I, w \cdot \lambda_i \geq a_i\}$, for finite I, and with each $\lambda_i \in \mathbb{R}^n$ and $a_i \in \mathbb{R}$. Consider any non-zero vector u in G_I. Then, u is rescale-optimal in G_I if and only if there exists $\mu \in \mathbb{R}^n$ that agrees on signs with u such that $\mu \cdot u = 1$ and μ is a positive linear combination of elements of $\{\lambda_i : i \in J_u\}$, where $J_u = \{i \in I : \lambda_i \cdot u = a_i\}$.*

Note that this implies that if non-zero u is rescale-optimal in G_I then J_u is non-empty, since $\mathbf{0}$ is the only positive linear combination of the empty set, and $\mu \neq \mathbf{0}$.

The proof uses the following lemmas. We first give a property that follows easily from standard results about convex cones.

Lemma 6 *Let Λ be a finite subset of $\mathbb{R}^n$ and let $\mu \in \mathbb{R}^n$. Then $\Lambda^* \subseteq (\{\mu\})^*$ if and only if $\mu \in co(\Lambda)$.*

Lemma 7 *Consider non-zero $u \in G_I$ (as defined above). Then u is rescale-optimal in G_I if and only if u is rescale-optimal in $G_{J_u} = \{w \in \mathbb{R}^n : \forall i \in J_u, w \cdot \lambda_i \geq a_i\}$.*

Lemma 8 *$G_{J_u} + \{-u\}$ is equal to $\{\lambda_i : i \in J_u\}^*$.*

Proof of Theorem 5

First consider $\mu \in \mathbb{R}^n$ such that $\mu \cdot u = 1$. Then it can be seen that $\{w : w \cdot \mu \geq 1\} + \{-u\} = (\{\mu\})^*$. Also, $G_{J_u} \subseteq \{w : w \cdot \mu \geq 1\}$ if and only if $G_{J_u} + \{-u\} \subseteq (\{\mu\})^*$ $\iff$ $\{\lambda_i : i \in J_u\}^* \subseteq (\{\mu\})^*$, using Lemma 8, which, by Lemma 6, is if and only if, $\mu \in co(\{\lambda_i : i \in J_u\})$.

By Lemma 7, u is rescale-optimal in G_I if and only if u is rescale-optimal in G_{J_u}, which, by Theorem 3, is if and only if there exists $\mu \in \mathbb{R}^n$ agreeing on signs with u such that $\mu \cdot u = 1$ and $G_{J_u} \subseteq \{w : w \cdot \mu \geq 1\}$, i.e., $\mu \in co(\{\lambda_i : i \in J_u\})$, by the earlier argument. □

We have the following corollary (using the same notation), which shows that testing if u is rescale-optimal in G_I can be performed in polynomial time: by first checking that $u \in G_I$ (i.e., for all $i \in I$, $u \cdot \lambda_i \geq a_i$), and then testing if a set of inequalities has a solution, using a linear programming solver.

Corollary 2 *u is rescale-optimal in G_I if and only if $u \in G_I$ and there exists non-negative reals r_i for each $i \in J_u$, (i.e., $i \in I$ such that $\lambda_i \cdot u = a_i$) and vector $\tau \in \mathbb{R}^n$ with for all $j \in \{1, \ldots, n\}$, $\tau(j) \geq 1$, and $\tau(j)u(j) = \sum_{i \in J_u} r_i \lambda_i$.*

5.2 Computation of Inference

For finite subsets Λ, Θ of $\mathbb{R}^n$, and arbitrary $\alpha, \beta \in \mathbb{R}^n$, we would like to be able to determine if $\alpha \gg_{\Lambda,\Theta} \beta$. Now, $\alpha \not\gg_{\Lambda,\Theta} \beta$ if and only if there exists u that is rescale-optimal in $G(\Lambda, \Theta)$ such that $u \cdot (\beta - \alpha) \geq 0$. Labelling Λ as $\{\lambda_i : i \in I\}$ and Θ as $\{\theta_k : k \in K\}$, it follows, using Theorem 5, that $\alpha \not\gg_{\Lambda,\Theta} \beta$ if and only if there exists $u \in \mathbb{R}^n$ and $\mu \in \mathbb{R}^n$, non-negative reals r_i for each $i \in I$ and non-negative reals s_k for $k \in K$, such that

- $u \cdot (\beta - \alpha) \geq 0$;
- $\forall i \in I$, $u \cdot \lambda_i \geq 1$, and $[u \cdot \lambda_i = 1$ or $r_i = 0]$;
- $\forall k \in K$, $u \cdot \theta_k \geq 0$, and $[u \cdot \theta_k = 0$ or $s_k = 0]$;
- $\forall j = 1, \ldots, n$, $u(j) = 0 \iff \mu(j) = 0$, and $u(j) > 0 \iff \mu(j) > 0$; and
- $\mu = \sum_{i \in I} r_i \lambda_i + \sum_{k \in K} s_k \theta_k$.

6 Experimental Testing

The experiments make use of a subset of a year's worth of real ridesharing records, provided by a commercial ridesharing system *Carma* (see `http://carmacarpool.com/`). We base our experiments on 13 benchmarks derived from this data-set. Each ridesharing alternative has 7 features, representing different aspects of a possible choice of match for a given user. Each benchmark corresponds to the inferred preferences of a different user. An input preference of alternative α_i over β_i leads to $\alpha_i - \beta_i$ being included in Λ. However, a pre-processing phase deletes some elements of Λ, in order to make it consistent (i.e., $\Lambda^> \neq \emptyset$), since in this paper we assume consistent preferences. (We assume no additional constraints, so $\Theta = \emptyset$.) More information about the data can be found in [Montazery and Wilson, 2016].

We randomly generate 100 pairs of (α, β), based on a uniform distribution for each feature. A pair (α, β) is called *decisive* for preference relation $\gg_\Lambda$ if either $\alpha \gg_\Lambda \beta$ or $\beta \gg_\Lambda \alpha$ hold, i.e., if α and β are comparable with respect to $\gg_\Lambda$; similarly, for $\gg^*_\Lambda$. (All 100 pairs turn out to be decisive for $>^{mm}_\Lambda$, as one would expect.) Table 1 shows the percentage of decisive pairs for $\gg_\Lambda$ and $\gg^*_\Lambda$, as well as the running time per request. CPLEX 12.6.2 is used as the solver on a computer facilitated by a Core i7 2.60 GHz processor and 8 GB RAM memory. Testing $\alpha \gg_\Lambda \beta$ is performed using quite a simple CPLEX model based on the approach in Section 5.2. Determining $\alpha \gg^*_\Lambda \beta$ is based on consistency of a set of linear constraints, so is fast.

The results indicate that for these benchmarks, $\gg_\Lambda$ is much more decisive than $\gg^*_\Lambda$. At the same time, $\gg_\Lambda$ is not equal to the maximum margin relation $>^{mm}_\Lambda$, so, in each case, rescaling makes a difference. Testing preference with respect to $\gg_\Lambda$ can be performed in reasonable time, the slowest instance being Benchmark 13, with a mean query time of around 5.3 seconds, based on 134 input preferences in Λ.

Table 1: A comparison, using 13 benchmarks, between preference relations $\gg_\Lambda$ and $\gg^*_\Lambda$.

	$\lvert\Lambda\rvert$	Decisive Pairs (%)		Time (msec)	
		$\gg_\Lambda$	$\gg^*_\Lambda$	$\gg_\Lambda$	$\gg^*_\Lambda$
1.	24	19	2	248	9
2.	29	94	0	1554	7
3.	31	18	0	421	6
4.	36	83	27	2478	6
5.	38	35	2	2123	5
6.	41	63	14	3006	5
7.	53	41	15	873	3
8.	55	97	11	1347	5
9.	62	54	1	1218	4
10.	94	62	6	2810	4
11.	127	67	9	3495	5
12.	129	77	0	1652	3
13.	134	68	19	5330	4
Avg	66	60	8	2217	5

7 Discussion

We have described a novel way of inducing a preference relation—by considering all possible rescalings when applying a linear SVM-based approach for preference learning—and derived formal results that allow its computation. Our experimental results indicate that the relation can be computed in a reasonable time for significantly sized instances, and that the relation can be considerably different from both the maximum margin relation and a simple cone-based relation. The results are also very relevant for a more general situation where one had restrictions on a set of allowable rescalings. There are a number of directions for further work, including: extensions to the case of inconsistent input sets and to the computation of which alternatives among a set can be optimal with respect to some rescaling; analysis—along similar lines as in this paper—to the other natural forms of invariance of feature spaces. Finally, it would be interesting to consider the application and generalisation of our results to other convex optimisation problems.

Acknowledgments

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289. We're very grateful for Carma for the use of their dataset. Thanks to the reviewers for their comments, which helped improve the final version of the paper.

References

[Aizerman *et al.*, 1964] A. Aizerman, E. M. Braverman, and L. I. Rozoner. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, 25:821–837, 1964.

[Ben-Hur and Weston, 2010] Asa Ben-Hur and Jason Weston. A users guide to support vector machines. In Oliviero

Carugo and Frank Eisenhaber, editors, *Data Mining Techniques for the Life Sciences*, volume 609 of *Methods in Molecular Biology*, pages 223–239. Humana Press, 2010.

[Brafman and Domshlak, 2009] Ronen I. Brafman and Carmel Domshlak. Preference handling - an introductory tutorial. *AI Magazine*, 30(1):58–86, 2009.

[Brafman, 2008] Ronen I Brafman. Preferences, planning and control. In *KR*, pages 2–5, 2008.

[Burges, 1998] Christopher JC Burges. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2):121–167, 1998.

[Fürnkranz and Hüllermeier, 2010] Johannes Fürnkranz and Eyke Hüllermeier. *Preference learning*. Springer, 2010.

[Greco *et al.*, 2010] Salvatore Greco, Vincent Mousseau, and Roman Slowinski. Multiple criteria sorting with a set of additive value functions. *European Journal of Operational Research*, 207(3):1455–1470, 2010.

[Herbrich *et al.*, 1999] Ralf Herbrich, Thore Graepel, and Klaus Obermayer. Support vector learning for ordinal regression. In *Artificial Neural Networks, 1999. ICANN 99. Ninth International Conference on (Conf. Publ. No. 470)*, volume 1, pages 97–102. IET, 1999.

[Joachims, 2002] Thorsten Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 133–142. ACM, 2002.

[Kazawa *et al.*, 2005] Hideto Kazawa, Tsutomu Hirao, and Eisaku Maeda. Order svm: a kernel method for order learning based on generalized order statistics. *Systems and Computers in Japan*, 36(1):35–43, 2005.

[Marinescu *et al.*, 2012] Radu Marinescu, Abdul Razak, and Nic Wilson. Multi-objective influence diagrams. In *Uncertainty in Artificial Intelligence (UAI)*, pages 574–583, 2012.

[Marinescu *et al.*, 2013] Radu Marinescu, Abdul Razak, and Nic Wilson. Multi-objective constraint optimization with tradeoffs. In *Proc. CP-2013*, pages 497–512, 2013.

[Montazery and Wilson, 2016] Mojtaba Montazery and Nic Wilson. Learning user preferences in matching for ridesharing. In *Proceedings of the 8th International Conference on Agents and Artificial Intelligence (ICAART 2016)*, volume 2, pages 63–73, 2016.

[Stolcke *et al.*, 2008] Andreas Stolcke, Sachin Kajarekar, and Luciana Ferrer. Nonparametric feature normalization for svm-based speaker verification. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 1577–1580. IEEE, 2008.

[Trabelsi *et al.*, 2011] Walid Trabelsi, Nic Wilson, Derek Bridge, and Francesco Ricci. Preference dominance reasoning for conversational recommender systems: a comparison between a comparative preferences and a sum of weights approach. *International Journal on Artificial Intelligence Tools*, 20(4):591–616, 2011.

[Wilson and Montazery, 2016] Nic Wilson and Mojtaba Montazery. *Preference Inference Through Rescaling Preference Learning (extended version of current paper including proofs)*. Available at http://ucc.insight-centre.org/nwilson/RescalingProofs.pdf, 2016.

[Wilson *et al.*, 2015] Nic Wilson, Anne-Marie George, and Barry O'Sullivan. Computation and complexity of preference inference based on hierarchical models. In *Proc. IJCAI-2015*, 2015.

[Wilson, 2014] Nic Wilson. Preference inference based on lexicographic models. In *Proc. ECAI-2014*, pages 921–926, 2014.